in: Proc. IEEE Int. Conf. on Systems, Man and Cybernetics (IEEE-SMC 2001), Tucson, pp. 672-677, IEEE Press, 2001

A CONTRIBUTION TO VISION-BASED LOCALIZATION, TRACKING AND NAVIGATION METHODS FOR AN INTERACTIVE MOBILE SERVICE-ROBOT

HORST-MICHAEL GROSS, HANS-JOACHIM BOEHME and TORSTEN WILHELM

Ilmenau Technical University, Department of Neuroinformatics
98684 Ilmenau, Germany
Horst-Michael.Gross@informatik.tu-ilmenau.de

Abstract

The paper presents vision-based robot navigation and user localization techniques of our long-term research project PERSES (PERsonal SErvice System), which aims to develop an interactive mobile shopping assistant that allows a continuous and intuitively understandable interaction with customers in a home store. Against this background, the paper describes a number of new or improved approaches, addressing challenges arising from the characteristics of the operation area, and from the need to continuously interact with users in a complex environment. With our approaches to vision-based or visually-controlled map building, self-localization and navigation as well as user localization and tracking, we want to make a contribution to the real-world suitability of interactive mobile service-robots in non-trivial application areas and demanding human-robot interaction scenarios.

Keywords

Service robots, human-robot systems, navigation, self-localization, person detection, visual tracking

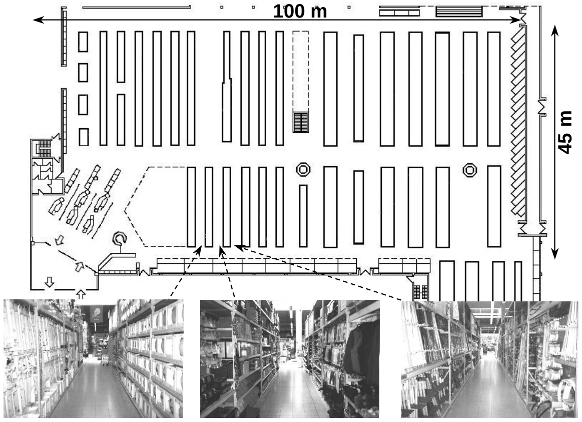

Figure 1: *(Top) Location plan of the home store, the experimental area of the PERSES-project. The topology of the store is characterized by many similar, long hallways of equal width (3 main passages of a length of about 100 meters and approx. 20 secondary hallways of 45 to 60 meters length). Because of their regular structure, the most of the hallways can be distinguished only visually by means of color or texture features; (Bottom) exemplary views of three hallways.*

1 Introduction

An interactive mobile service robot, e.g., a shopping assistant, should be able to actively observe its operation area, to detect, localize, and contact potential users, to interact with them continuously, and to adequately offer its specific services. Typical service tasks we want to solve in the PERSES project are to guide the user to desired areas or articles within a home store *(guidance function)* or to follow him as a user-specific mobile information kiosk while continuously observing the user and his behavior *(companion function)* (see [6], too). In the context of this scenario, the following interaction and navigation tasks had to be tackled: (a) building and maintaining large-scale maps as well as continuous self-localization of the robot in the operation area, (b) robust avoidance of static and dynamic obstacles during navigation, (c) navigation to desired market areas acting as a guide, (d) visual user localization within a pre-defined operation area, (e) acoustic localization of a potential user clapping his hands or shouting a command to attract attention, (f) fast learning of a visual model of the current user and online adaptation of that model due to the varying appearance of the user in the course of the shopping process, (g) robust visual user tracking both while standing still and during self-movement of the robot, (h) recognition of simple spoken commands, and, for the future, (i) recognition of gesticulated user instructions as a kind of non-verbal communication.

Up to now, in many mobile robot applications the robots perceive their surroundings mainly by means of distance sensors (laser, sonar). To accommodate the challenges that arise from the specificity of our interaction-oriented scenario and the characteristics of the operation area, a highly unstructured, dynamic and crowded environment, we placed special emphasis on vision-based methods both for human-robot interac-

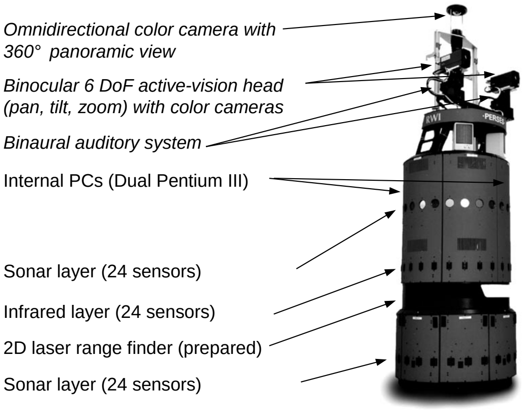
Omnidirectional color camera with 360° panoramic view
Binocular 6 DoF active-vision head (pan, tilt, zoom) with color cameras
Binaural auditory system
Internal PCs (Dual Pentium III)
Sonar layer (24 sensors)
Infrared layer (24 sensors)
2D laser range finder (prepared)
Sonar layer (24 sensors)
RWI
PERSES

Figure 2: *Experimental platform PERSES, an extended version of a B21 robot. In addition to the standard equipment of two sonar and one IR-layers, PERSES is equipped with (i) an omnidirectional color camera with a* 360^o *panoramic view used for user localization and tracking, self-localization and obstacle avoidance, (ii) a binocular 6 DoF active-vision head with 2 frontally aligned color cameras used for user verification and tracking and odometry correction, and (iii) a binaural auditory system for acoustic user localization.*

tion and robot navigation. The operation area, for example, is characterized by many similar, long hallways of equal width and a great number of critical obstacle configurations, e.g., objects hanging down from the ceiling or jutting out of shelves, lost shopping carts, etc. Many of the obstacles cannot be perceived reliably by distance sensors which operate in certain planes in 3D space. Moreover, self-localization methods using distance sensors can produce numerous ambiguities preventing a quick and reliable self-localization or relocalization of the robot. In contrast, vision-based approaches do not show these limitations, but supply a much greater wealth of information about the structure of the local surroundings and the current behavior of the interaction partner. This is not only required for an efficient and safe navigation in this critical environment, but also for a stable contact between robot and user, which, in turn, is a fundamental prerequisite for a continuous interaction in the course of the shopping process. Fig. 1 is to illustrate some of the challenges of the operation area, a typical home store located in the capital of Thuringia. The robot PERSES we use as experimental platform is a standard B21 robot additionally equipped with color cameras for interaction and navigation (Fig. 2).

In the following, we present a number of new or improved approaches, addressing challenges arising from the characteristics of the operation area, and from the need to continuously interact with users in a complex environment: a) visually-controlled building and maintaining of large scale probabilistic occupancy maps of the operation area, b) vision-based robot self-localization that combines panoramic views of the robot's omidirectional color camera with the Monte Carlo Localization (MCL) [5], c) vision-based obstacle avoidance using optical flow fields and inverse perspective mappings of the panoramic image, and d) vision-based user localization and tracking by means of specific cues and particle filters [7]. Fig. 3 illustrates the essential subtasks of the interaction and navigation process from a functional point of view. All vision-based methods presented in this paper are highlighted by gray background.

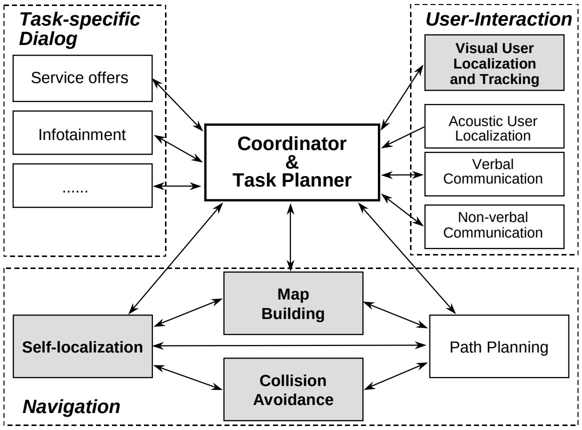

Figure 3: *Overview of the project-specific subtasks of the interaction and navigation process from a functional point of view.*

2 Vision-based navigation and interaction methods

2.1 Building and maintaining a global map

To navigate reliably in indoor environments, a robot must know where it is. This includes both the ability of globally localizing the robot from scratch, as well as tracking the robot's position once its location is known. In PERSES, we use two types of maps for self-localization and navigation: (i) grid-based occupancy maps and (ii) a grid of panoramic views of local surroundings (see section 2.2). The maps are learned from sensor data (sonar scans, odometry readings and panoramic images) collected when manually joy-sticking the robot through the store. Up to now, the building of the local and global occupancy maps for navigation and path planning is based on sonar data and odometry readings. One major problem using odometry data is their increasing error over time, especially concerning the orientation angle. This problem is well known and leads to the fact that a global map generated along a closed-loop course cannot be really closed without additional efforts (see Fig. 5, left). To attenuate this effect, we utilize a specific feature of our market floor (ground) which shows a rectangular structure caused by tiles that are uniquely oriented

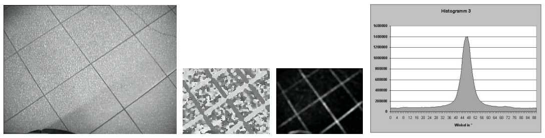

Figure 4: *General idea of our vision-based odometry correction considering a specific feature of the market floor: a) image of the floor in front of the robot, b) local orientation tensors (orientations are coded as gray values), c) confidences of local orientations (low-black, high-white), d) histogram of confidence-weighted local orientations, the dominant orientation (center of gravity) is a significant measure for the accurate orientation of the robot in the interval* $0^o - 90^o$

across the whole market area. The idea is illustrated in Fig. 4: a top-down oriented on-board camera acquires images of the floor in front of the robot. By continuously estimating the dominant orientations within these images, we can calculate the accurate orientation of the robot and, therefore, substitute the rotation angle supplied by odometry by the orientation determined visually. Hence, it is possible to eliminate the orientation error, and subsequently, the position error. Under the assumption that the initial position and orientation of the robot are known, this method allows an accurate, iterative position tracking as required for map building. Fig. 5 and 6 illustrate the efficiency of this specific method for vision-based odometry correction for map building. Fig. 5 shows the resulting occupancy maps of a section of the store (60 by 20 meters) without (left) and with (right) odometry correction, while Fig. 6 presents an occupancy map of the complete store (100 by 60 meters). Of course, the proposed approach does not hold in a more general framework, but is very well suited for our specific environment. The introduced method for vision-based odometry correction allows to efficiently build very large and exact occupancy maps on the fly - without the computationally expensive EM-algorithm for localization estimation and map building [2], which is extensively exploited for mobile robot applications, e.g., in [9].

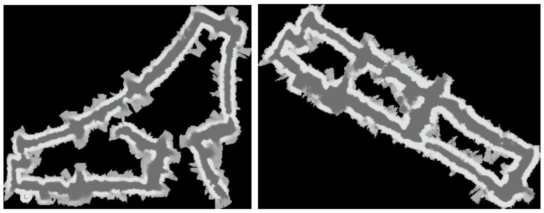

Figure 5: *Results of the occupancy map building: sonar-based maps of a store section (60 by 20 meters; path length: 250 m). Gray-values code occupancy probabilities: white - occupied (obstacle), gray - free-space, black - unexplored. (Left) without vision-based odometry correction: the closed-loop course cannot be closed, because the error of the odometry-based estimation of the rotation angle finally amounts to* 90^o*; (Right) with vison-based odometry correction, now the closed-loop course can be closed exactly.*

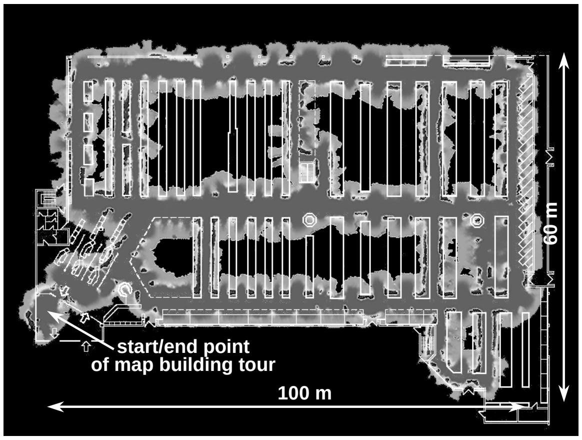

Figure 6: *Result of visually corrected building of an occupancy map of the home store. The map was learned from sonar scans collected when manually joy-sticking the robot through the three main passages and a number of secondary hallways (total path length: 650 meters). In order to demonstrate the accuracy of the map building, the exact location map (white CAD-map) and the occupancy map created on the fly are overlaid. The final localization error at the start/end point of the tour located in the entrance area is lower than 0.5 meter.*

2.2 Vision-based robot self-localization

As mentioned above, the topology of the store area is characterized by many similar, long hallways of equal width (see Fig. 1). For this reason, self-localization methods based on distance sensors can produce numerous ambiguities preventing a quick self-localization and relocalization in case of a complete loss of positioning. Because the visual input from the omnidirectional color camera supplies a much greater wealth of information about the structure of the local surroundings, we expect to defuse that problem and to accelerate relocalization significantly. Therefore, we currently develop an approach for vision-based self-localization that combines panoramic views of the omni-camera with the Monte Carlo Localization (MCL) developed by FOX [5]. MCL is a relatively new algorithm for robust and efficient self-localization of mobile robots. It is a version of Markov localization [8, 4], a family of probabilistic approaches for approximating a multi-modal probability distribution coding the robot's belief $Bel(l)$ for being at position $l = (x, y, \varphi)$ in the state space of the robot. x and y are the robot's coordinates in a world-centered Cartesian reference frame, and φ is the robot's orientation. MCL applies sampling techniques to represent the posterior belief $Bel(l)$ for being at position l by a set of N weighted, random samples S. Samples in MCL are of the type $\langle\langle x, y, \varphi\rangle, p\rangle$, where $p \geq 0$ is a numerical

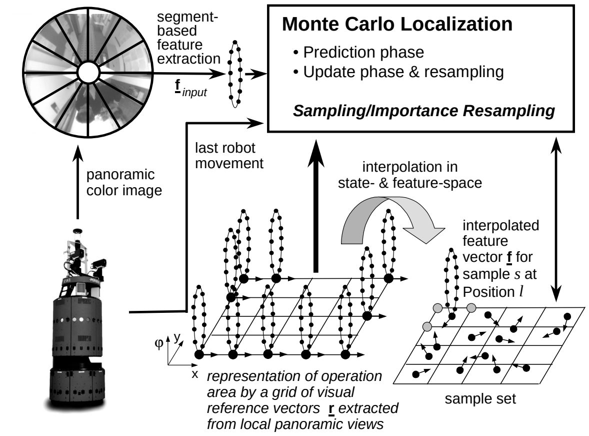
segment-based feature extraction
$\mathbf{f}_{input}$
Monte Carlo Localization
• Prediction phase
• Update phase & resampling
Sampling/Importance Resampling
last robot movement
panoramic color image
interpolation in state- & feature-space
interpolated feature vector f for sample s at Position l
φ
y
x
representation of operation area by a grid of visual reference vectors r extracted from local panoramic views
sample set

Figure 7: *General idea of our view-based Monte Carlo Localization. The approach is based on a grid-based representation of the operation area by a set of panoramic views of local surroundings. The visual features (in the simplest case, mean RGB color values) are extracted from annulus segments of the omnidirectional color image. For more details see [6].*

weighting factor, analogous to a discrete probability. Because the sample set constitutes a discrete approximation of the continuous probability distribution, the MCL approach is computationally efficient, it places computation just “where needed”. Additionally, it is more accurate than Markov localization with a fixed cell size, as state represented in samples is not discretized [5]. This allows a self-localization and position tracking with sub-grid accuracy. In analogy with the MCL algorithm presented in [5], our view-based MCL (see Fig. 7) proceeds in two phases, the *Prediction phase* (robot motion) and the *Update phase* which are described more in detail in [6]. Fig. 8 illustrates a typical result of a view-based self-localization and position tracking experiments recently executed in a large section of the store ($28 \times 17m^2$). Despite the uniformity of the learned three hallways (see Fig. 1) and the coarse grid-space of $60\ cm$, the view-based MCL yields very precise localization results already after a few robot movements. After 5 to 7 movements (about 10 meters), the difference between estimated and correct position of the robot was not larger than 20% of the grid-space, i.e., about 10 cm. In earlier experiments presented in [6] we investigated the influence of occlusions on the localization accuracy. In these experiments, a few people were standing around the robot at a very low distance and occluded large regions of the panoramic image (40-50%). Despite this occlusion, the MCL still generated good localization results. In this case, the difference between estimated and correct position was not larger than the grid space. The results of the empirical experiments confirm the robustness of this vision-based self-localization method, however, it still has to demonstrate its capabilities in still larger and more complex operation areas of the home store.

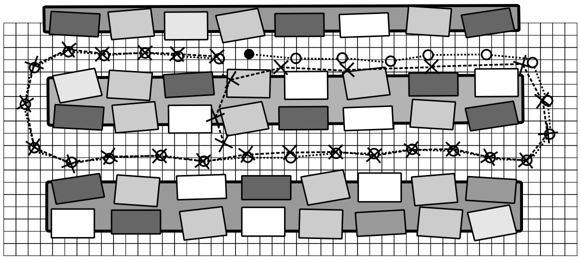

Figure 8: *Self-localization and tracking experiment executed in a large section of the store ($28 \times 17m^2$) covering 3 hallways and 4 long shelves presented in Fig. 1 (not completely shown here). The grid-space is 60 cm, the total path length is 50 m. The correct positions of the robot are marked by white circles ○, the estimated positions (centers of gravity of the samples with highest weighting factors) by ×. Because the localization error is lower than 20% of the grid-space (about 10 cm), in most cases the ○ and × are overlapping.*

2.3 Vision-based obstacle avoidance

Before navigation commands (e.g., from the path planning subsystem) are executed, they are passed to the *obstacle avoidance* module (see Fig. 3) which suppresses commands impossible according to the current obstacle configuration in front of the robot. In addition to scan-based obstacle avoidance methods, we currently investigate vision-based methods. The need for supplemental vision-based methods for obstacle avoidance is due to numerous obstacles that cannot be perceived reliably by 2D distance sensors (sonar, laser) because of their specific form, size or height (e.g., objects jutting out of shelves or hanging down from the ceiling). Against this background, local navigation methods based on optical flow fields and inverse perspective mappings are currently investigated in our lab. Based on the images provided by the omnidirectional camera, we estimate panoramic optical flow fields (Fig. 9) that are used for an efficient bee-like local navigation. The underlying control mechanism [3] tries to balance the optical flow in both hemispheres of the visual field, which results in a collision-free locomotion in the middle of obstacle configurations (e.g., hallways). This approach, however, still has a number of insufficiencies in free-space and dead-ends.

2.4 Visual user localization and tracking

One of the major problems of this scenario consists in the robust localization, verification and tracking of a user in the area. A detailed explanation of our multi-cue approach is given in [1].

User localization: The visual user localization performs a motion-based foreground-background segmentation in the image sequence provided by the omnidirectional camera, and returns the angle to the cen-

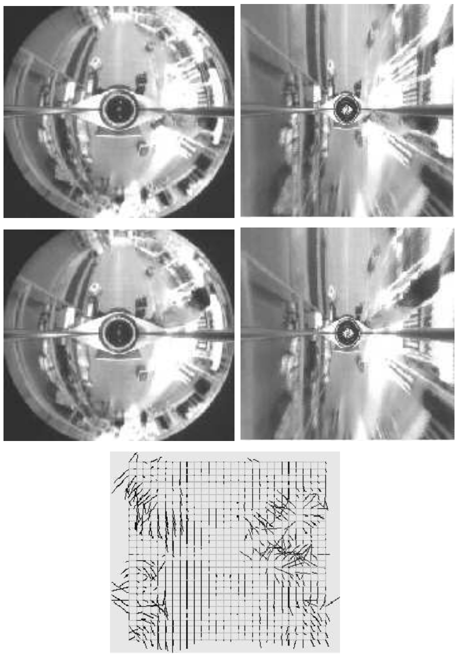

Figure 9: *Results of inverse perspective mappings of successive omnidirectional views of the local surroundings (left) onto virtual top-views (right) that are used to estimate a panoramic optical flow field (bottom). It is obvious that the person moving from the positions at 3 to 2 o'clock in the omni- and top-view images causes large flow vectors, while the static objects on the right just cause small ones. This technique simplifies the detection of dynamic obstacles during self-movement of the robot. As a consequence of the inverse perspective mapping, towering objects or objects jutting out of shelves produce larger flow vectors than objects located near the ground. The flow vectors on the left side in the image are the result of the self-movement of the robot near the shelves.*

ter of gravity of the closest region moving towards the robot. In the waiting position or while standing still, this motion-based segmentation determines candidate regions that indicate if and where potential users could be in the surroundings of the robot.

User verification: The verification of the localization hypothesis is triggered, when the robot was turned to the moving object. Due to this turn, a potential user should be localized in front of the robot, allowing to observe him by the frontally aligned cameras as well as by the omnidirectional camera. To evaluate if there really is a person and if the person could be willing to interact with the robot, we developed a verification system that integrates different visual cues. Person verification should highlight that regions most likely cover the upper part of a person. Fig. 10 gives an overview of our task-specific multi-cue approach that integrates the following visual cues: facial structure, head-shoulder-contour and skin color. Because we want to localize people even at different distances from the robot, a multiresolution pyramid (scale space with five fine-to-coarse resolutions) transforms the images into a multi-scale representation. The cue modules sensitive to facial structure and head-shoulder contour operate at all levels of the grayscale pyramid, while the cue module for skin color detection uses the original color image. After superposition of the corresponding feature maps, a 3D-Winner-Take-All process within the saliency pyramid selects that region most likely covering the upper part of a person. Fig. 11 shows typical verification results obtained in the highly structured environment of the home store. By this multi-cue approach, we can handle varying environmental circumstances much easier, which, for instance, can make the skin color detection difficult or almost impossible.

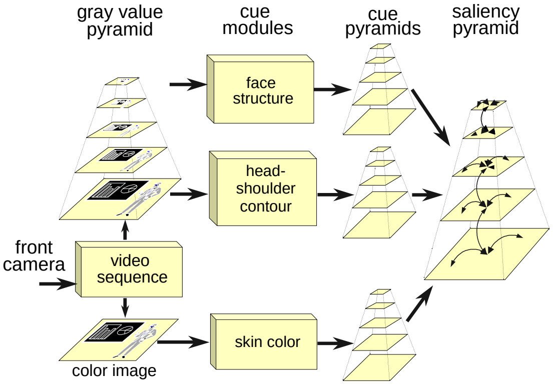

Figure 10: *Multiple-cue approach for user verification.*

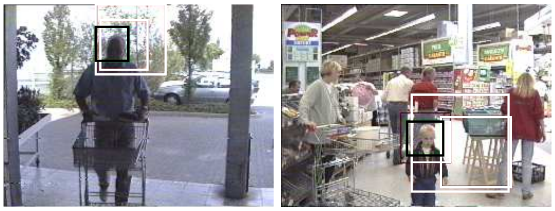

Figure 11: *Typical verification results for different situations in the home store; size of the frames corresponds to the respective level of the scale space - small frames correspond to levels of high-resolution and vice versa. Final localization results are marked as black frames. (Left) back light scene taken in the entrance area, where only the contour detection can provide a confident contribution for verification; (Right) crowded area in the store - the child is selected as final localization result because it is the only subject that fulfills all 3 criteria of the multi-cue approach: face and upper part of the body are oriented frontally towards the robot, skin color can be detected clearly.*

User tracking: During the whole interaction cycle, the robot tries to keep continuous contact to the current customer via visual tracking. Person verification provides the initialization for this tracking process, and tracking can be done via the omnidirectional

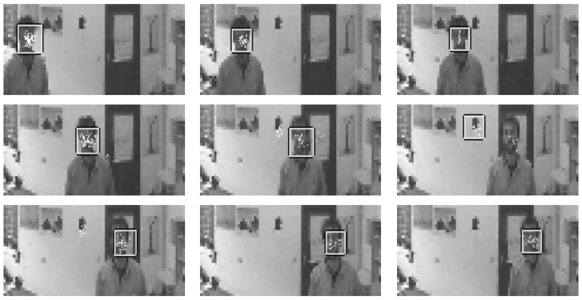

Figure 12: *User tracking in a sequence of panoramic images provided by the omnidirectional camera. In this sequence, only the frontal part ($+/-50^o$) of the panoramic visual field is log-polar transformed and analyzed. The particles with the highest probability (coding significant user localization hypotheses) are mapped onto the images as white dots. The black/white frames indicate those locations where the tracked user is most likely expected (center of gravity of the largest particle cloud).*

camera as well as via the frontally aligned cameras. The tracking procedure is based on the *Condensation* algorithm [7], widely accepted as a powerful and efficient method for tracking arbitrarily shaped probability distributions [5]. The features underlying the tracking process, a combination of head-shoulder contour detection and skin color modeling, which turned out to be appropriate, were derived from the user verification procedure. An exemplary tracking sequence is given in Fig. 12, where samples of a longer run of the tracking system are shown. To improve the robustness of the racking procedure, the person verification module runs in parallel and supplies further candidate points. As long as the contact to the current user can be continuously updated, no articulation of the robot is needed. If the robot detects a situation where the user is lost, it provides a speech output to ask the user to reduce the distance to the robot. Alternatively, the guidance to a desired market position is interrupted and the robot moves towards its current user to prevent losing contact.

3 Conclusions and Outlook

In the paper, special emphasis has been placed on vision-based methods for both map building, self-localization and local navigation of the mobile robot PERSES, and localization and tracking of potential users to accommodate the challenges that arise from the characteristics of the environment and from the need to continuously interact with people. For the already implemented subsystems, we presented both important aspects of the methodological background and results of ongoing experiments executed in the operation area, a home store. The experimental results are promising and illustrate the functionality, but also the still existing weaknesses of the already realized vision-based methods. Future research will concentrate on the integration of the methods presented here into a prototypical interaction cycle and on the implementation of a probabilistic control architecture that allows the self-organization of efficient strategies for multi-modal human-robot interaction during the shopping process. In order to model the current state of the user within the interaction cycle and to select the most effective interaction-strategy, the integration of procedures for speech recognition and interpretation of gestures and/or body language, and a verbal, mimic, or body language-based articulation of the robot back to the user is required. Such an intuitively understandable, bidirectional interaction between robot and user still remains a challenge to realize a user-friendly guidance and companion function.

Acknowledgement: The authors wish to thank the PERSES and MILVA mobile robot groups, particularly J. Key, A. Liebetrau, and A. Koenig for implementations, experiments and discussions.

References

[1] H.-J. Boehme et al. An Approach to Multimodal Human-Machine Interaction for Intelligent Service Robots. subm. to: *Eurobot 2001*, Lund

[2] A. Dempster, A. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39 (1977) 1-38

[3] A.P. Duchon, W.H. Warren, L.P. Kaelbling. Ecological Robotics. *Adaptive Behavior*, 6 (1998) 473-507

[4] D. Fox, W. Burgard, and S. Thrun. Markov Localization for Mobile Robots in Dynamic Environments. *Journ. of Artif. Intelligence Research*, 11 (1999) 391-427

[5] D. Fox, W. Burgard, F. Dellaert, S. Thrun. Monte Carlo Localization: Efficient Position Estimation for Mobile Robots. In: *Proc. AAAI-99*, 1999

[6] H.-M. Gross, H.-J. Boehme. PERSES - a Vision-based Interactive Mobile Shopping Assistant. in: *Proc. IEEE-SMC 2000*, pp. 80-85

[7] M. Isard, A. Blake. CONDENSATION - conditional density propagation for visual tracking. *Int. Journal on Computer Vision*, 29 (1): 5-28, 1998

[8] S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artif. Intell.*, 99 (1998) 21-71

[9] S. Thrun, W. Burgard, and D. Fox. A Probabilistic Approach to Concurrent Mapping and Localization for Mobile Robots. *Machine Learning and Autonomous Robots*, 31 (1998) 1-25